LECTURE 3 : BASIS FUNCTION

[I] Motivation : Polynomial Functions and Basis Functions

To model our data, in the previous discussion, we used polynomial functions:

$$y(x_n, \bar{w}) = \sum_{m=0}^{M} w_m x_n^m = \bar{w}^T \bar{x}_n$$

These functions are part of more general functions called Basis functions.

↓

Basis function models:

$$y(x_n, \bar{w}) = w_0 + \sum_{m=1}^{M-1} w_m \phi_m(x_n)$$

Defining $\phi_0(x_n) = 1$, we can write : $y(x_n, \bar{w}) = \sum_{m=0}^{M-1} w_m \phi_m(x_n) = \bar{w}^T \bar{\phi}(x_n)$

where $\phi_m()$ is called a basis function

↓

For polynomial functions : $\boxed{\phi_m(x_n) = x^m}$

Other basis functions :

The term $1/\sqrt{2\pi s^2}$ is not used since ϕ_m is not a probability function.

Radial Basis Functions ← [•] Gaussian basis function : $\boxed{\phi_m(x_n) = \exp\left(-\frac{(x_n - \mu_m)^2}{2s^2}\right)}$

[•] Sigmoid basis function: $\boxed{\phi_m(x_n) = \sigma\left(\frac{x_n - \mu_m}{s}\right)}$

where $\sigma(a) = \frac{1}{1 + \exp(-a)}$

Questions

① Why do we need other types of basis functions, and not just stick with polynomial functions?

② How do the other basis functions work ?

[2] How Basis Functions Work?

$$y(x_n, \bar{w}) = \sum_{m=0}^{M-1} w_m \phi_m (x_n)$$

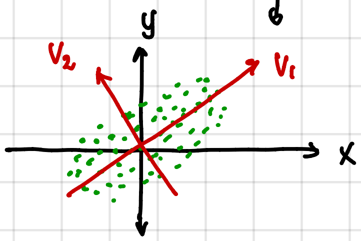
The above definition means we can represent any function using ϕ_m, basis functions.

↓

We can think basis functions as basic blocks, like lego blocks, where we use to represent any shape of objects.

↓

Example 1 :



We can consider $\bar{v}_1$ and $\bar{v}_2$ as the basis functions, because we can represent the location of the green dots with $\underbrace{\bar{v}_1 \text{ and } \bar{v}_2}$: $a\bar{v}_1 + b\bar{v}_2$

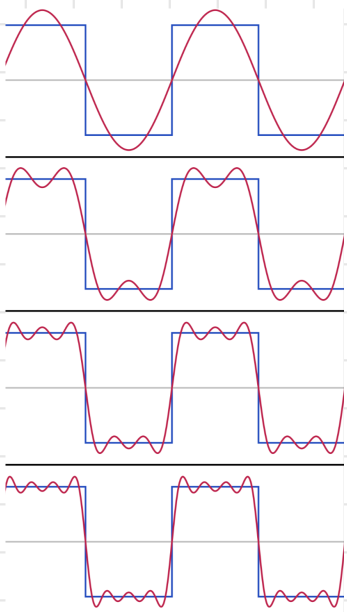
they are called principle components.

Example 2 :

Fourier series : it can represent any periodic functions as the sum of sine or/and cosines.

$$f(x) = \frac{1}{2} a_0 + \sum_{n=1}^{\infty} a_n \cos(nx) + \sum_{n=1}^{\infty} b_n \sin(nx)$$

$$= \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \sin(nx)$$



See: ① Wikipedia on Fourier series → there is a nice demo of Fourier series as basis functions.

② Spline functions : geometrie.foretnik.net/files/NURBS-en.swf

[3] Gaussian Basis Functions

$$\boxed{y(x_n, \bar{w}) = \sum_{m=0}^{M-1} w_m \phi_m(x_n)} \quad ; \quad \boxed{\phi_m(x_n) = \exp\left(-\frac{(x_n - \mu_m)^2}{2s^2}\right)}$$

$\Downarrow$

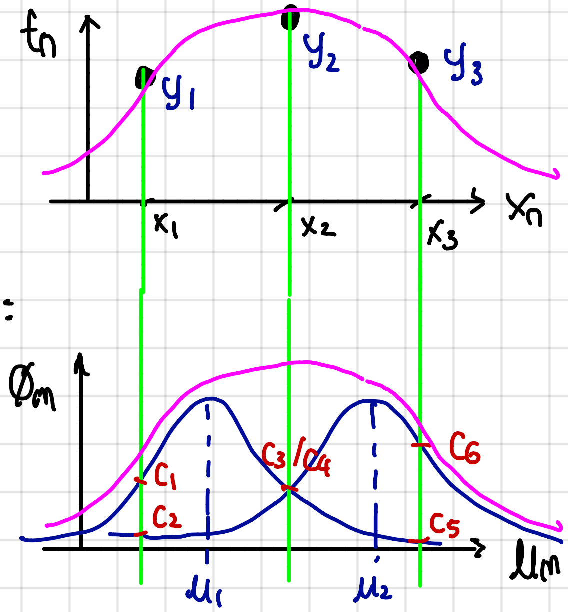
Two important parameters: μ_m and s

[•] Basic ideas :

1. Assume $M = 3$; $w_0 = 0$, $w_1 = 1$, $w_2 = 1$

and we have 3 points of data :

M=3 :



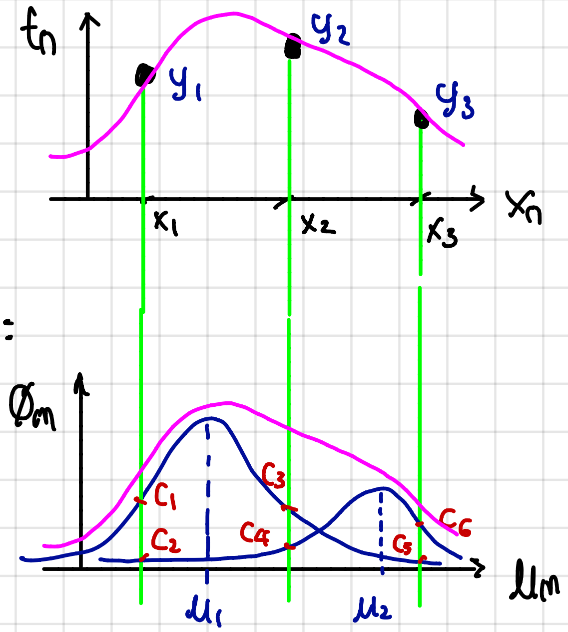
$$y(x_1, \bar{w}) = y_1 = \underset{C_1}{\phi_1(x_1)} + \underset{C_2}{\phi_2(x_1)}$$

$$y(x_2, \bar{w}) = y_2 = \underset{C_3}{\phi_1(x_2)} + \underset{C_4}{\phi_2(x_2)}$$

$$y(x_3, \bar{w}) = y_3 = \underset{C_5}{\phi_1(x_3)} + \underset{C_6}{\phi_2(x_3)}$$

2. For the above example, if we have $w_2 = 1/2$; it means the second Gaussian gets half smaller :

M=3 :



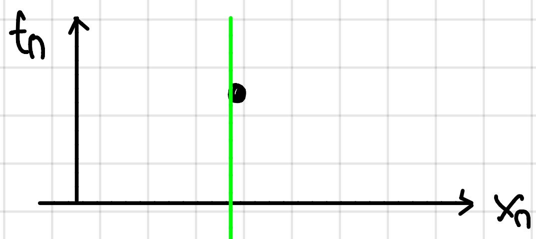
$$y(x_1, \bar{w}) = y_1 = \underset{C_1}{\phi_1(x_1)} + \underset{C_2}{\phi_2(x_1)/2}$$

$$y(x_2, \bar{w}) = y_2 = \underset{C_3}{\phi_1(x_2)} + \underset{C_4}{\phi_2(x_2)/2}$$

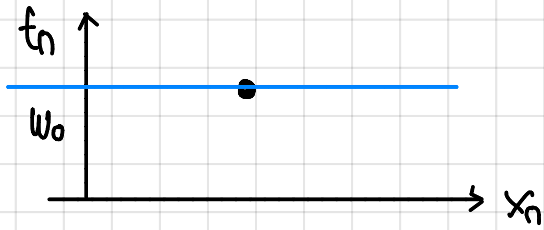
$$y(x_3, \bar{w}) = y_3 = \underset{C_5}{\phi_1(x_3)} + \underset{C_6}{\phi_2(x_3)/2}$$

[•] Other examples (when $\bar{\omega}$ is unknown)

1. For 1 point of data:

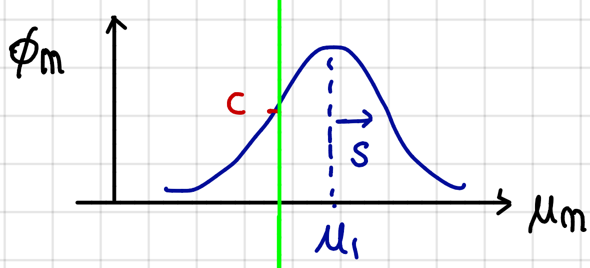


M=1:



$$y(w_0) = w_0$$

M=2:

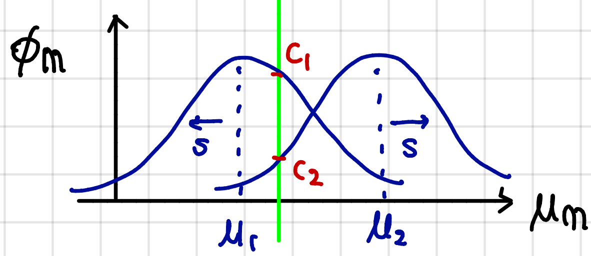


$$y(x_1, \bar{\omega}) = w_0 + w_1 \phi_1(x_1) = \underline{w_0 + w_1 c} \quad ; \quad c = \text{a constant}$$

one equation, 2 unknowns

(underdetermined system: $N < M$)

M=3:



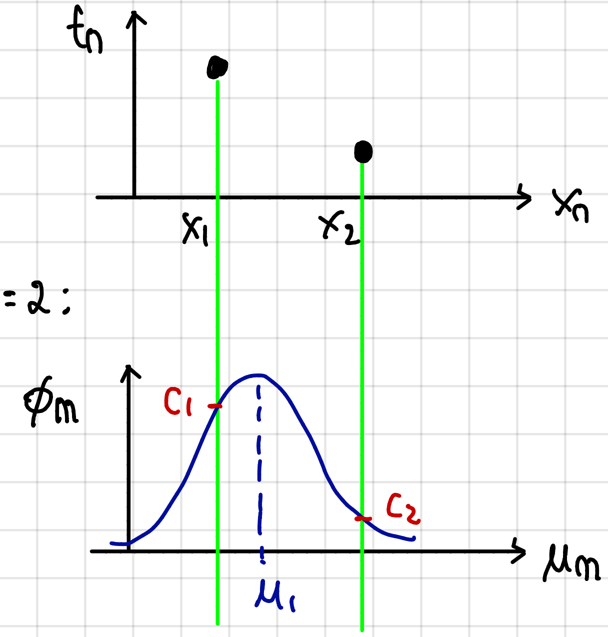
$$y(x_1, \bar{\omega}) = w_0 + w_1 \phi_1(x_1) + w_2 \phi_2(x_2)$$

$$= \underline{w_0 + w_1 c_1 + w_2 c_2} \quad ; \quad c_1, c_2 = \text{constants}$$

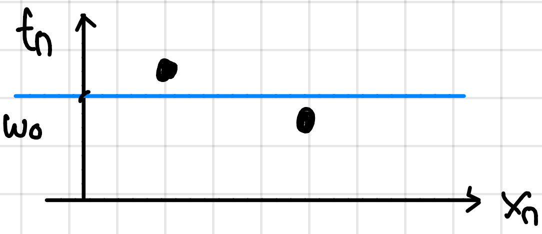
one equation, 3 unknowns

2. For 2 points of data:

M = 2:



M = 1 :

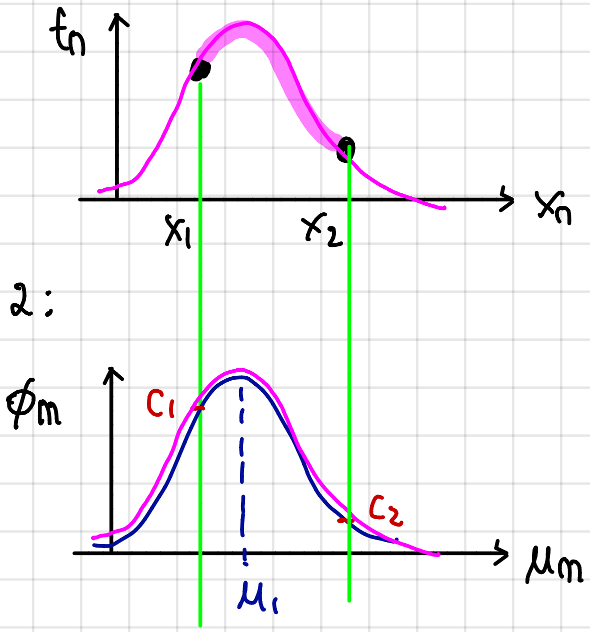


$y(w_0) = w_0$

$$\left.\begin{aligned} y(x_1,\bar{w}) &= w_0 + w_1\,\phi_1(x_1) \approx w_0 + w_1 c_1 \\ y(x_2,\bar{w}) &= w_0 + w_1\,\phi_1(x_2) \approx \underbrace{w_0 + w_1 c_2}_{\text{2 equations, 2 unknowns}} \end{aligned}\right\} \text{ We can obtain } w_0 \,\&\, w_1$$

Once, we know w_0 & w_1, we can apply the basis functions to all values of x_n. Assuming $w_0 = 0$ & $w_1 = 1$, then:

M = 2:



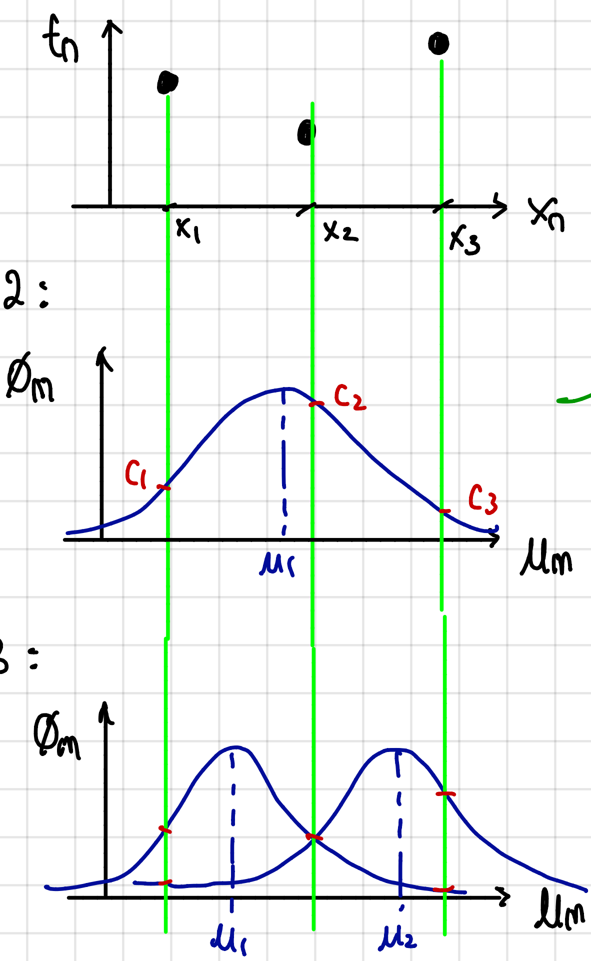
$y_1 = \phi_1(x_1)$

$y_2 = \phi_1(x_2)$

3. For 3 points of data:

M=2:

M=3:

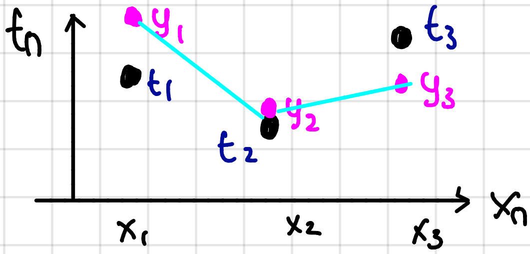


$$y(x_1,\bar{\omega}) = w_0 + w_1\,\phi_1(x_1)$$
$$y(x_2,\bar{\omega}) = w_0 + w_1\,\phi_1(x_2)$$
$$y(x_3,\bar{\omega}) = w_0 + w_1\,\phi_1(x_3)$$

$$\downarrow$$

$$y_1 = y(x_1,\bar{\omega}) = w_0 + w_1 c_1$$
$$y_2 = y(x_2,\bar{\omega}) = w_0 + w_1 c_2$$
$$y_3 = y(x_3,\bar{\omega}) = w_0 + w_1 c_3$$

over determined



Since the equations are over determined, once we obtained w_0 and w_1, we can't produce exact values of t_1, t_2 & t_3.

$$y(x_1,\bar{\omega}) = w_0 + w_1\,\phi_1(x_1) + w_2\,\phi_2(x_1)$$
$$y(x_2,\bar{\omega}) = w_0 + w_1\,\phi_1(x_2) + w_2\,\phi_2(x_2)$$
$$y(x_3,\bar{\omega}) = w_0 + w_1\,\phi_1(x_3) + w_2\,\phi_2(x_3)$$

These are non-linear functions

For polynomials: $\phi_1(x) = x$; $\phi_2(x) = x^2$

The result of M=3 will fit more on the points.

Q: Why polynomial function can be unstable numerically?

A: Imagine $M = 51$, where we need to estimate $\bar{\omega} = (w_0, \ldots, w_{50})$ from the following set of equations:

$$
\begin{aligned}
y_1 &= w_0 + w_1 x_1 + w_2 x_1^2 + \cdots + w_{50} x_1^{50} \\
y_2 &= w_0 + w_1 x_2 + w_2 x_2^2 + \cdots + w_{50} x_2^{50} \\
&\vdots \qquad\qquad \vdots \\
y_N &= w_0 + w_1 x_N + w_2 x_N^2 + \cdots + w_{50} x_N^{50}
\end{aligned}
$$

If any x_i is large ($x_i = 20$), then x_i^{50} is a huge number. Assuming the underlying curve is simple (e.g. quadratic), many w's (particularly those of higher degrees of polynomial) will be very-very small (very close to zero). These two factors of huge numbers of x_i^m and w's can cause numerical problems in the implementation.

Q: Any other advantages of using Radial Basis Functions (e.g. Gaussian basis)?

A: (1) It's easy to process a high dimensionality of $\underset{D\times 1}{\bar{x}_i}$ (e.g. D)

$$
\begin{aligned}
y(\bar{x}_n, \bar{w}) &= \sum_{m=0}^{M-1} w_m \phi_m(\bar{x}_n) \\
&= \underset{1\times M}{\bar{\phi}^T}(\underset{D\times 1}{\bar{x}_n})\,\underset{M\times 1}{\bar{w}} = \underset{1\times M}{\bar{\phi}_n^T}\,\underset{M\times 1}{\bar{w}}
\end{aligned}
$$

(2) We can express $\underset{D\times 1}{\bar{x}_n}$ using $\underset{M\times 1}{\bar{\phi}_n}$, which implies that we transform $\bar{x}_n$ using $\bar{\phi}_n$ in a such a way the dimensionality changes from $D\times 1$ to $M\times 1$.

• MLE using Basis Functions

Assuming : $t_n = y(x_n, \bar{w}) + \varepsilon$; ε = Gaussian noise

$$p(t_n \mid x_n, \bar{w}, \beta) = G(t_n ; y(x_n, \bar{w}), \beta)$$

Given input data $\bar{X} = \{x_1, x_2, \dots, x_N\}$ that are independent to each other, then :

the likelihood → $$p(\bar{t} \mid \bar{x}, \bar{w}, \beta) = \prod_{n=1}^{N} G(t_n ; \underbrace{\bar{w}^T \bar{\phi}(x_n)}_{\text{Basis functions}}, \beta)$$

If x_n is in $\mathbb{R}^D$, meaning x_n is a vector occupying a D dimensional space, then $\underset{D\times N}{X} = \{\underset{D\times 1}{\bar{x}_1}, \underset{D\times 1}{\bar{x}_2}, \dots, \underset{D\times 1}{\bar{x}_N}\}$. Consequently, the likelihood becomes:

$$p(\bar{t} \mid X, \bar{w}, \beta) = \prod_{n=1}^{N} G(t_n ; \bar{w}^T \phi(x_n), \beta)$$

$$\Downarrow$$

Recall that MLE for the curve fitting is wri n as:

$$\{\bar{w}, \beta\}^* = \underset{\{\bar{w}\}\{\beta\}}{\operatorname{argmax}} \; p(\bar{t} \mid X, \bar{w}, \beta)$$

$$= \underset{\{\bar{w}\}\{\beta\}}{\operatorname{argmax}} \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\beta^{-1}}} \exp\left(-\frac{\beta}{2} \; n - \bar{w}\bar{\phi}(\bar{x}_n))^2\right)$$

$$= \underset{\{\bar{w}\}\{\beta\}}{\operatorname{argmin}} \sum_n \frac{1}{2} \log 2\pi - \sum_n \frac{1}{2} \log \beta + \sum_n \frac{\beta}{2} \left(t_n - \bar{w}\bar{\phi}(\bar{x}_n)\right)^2$$

Optimization:

For $\bar{w}$: $$\underset{M\times 1}{\frac{\partial}{\partial \bar{w}}} \sum_n \frac{\beta}{2} \left(t_n - \bar{w}\phi(\bar{x}_n)\right)^2 = \boxed{\underset{M\times 1}{\beta \sum_n \left(\underset{1\times 1}{t_n} - \underset{1\times M}{\bar{w}^T} \underset{M\times 1}{\bar{\phi}(\bar{x}_n)}\right) \underset{1\times M}{\bar{\phi}(x_n)}} = 0} = \underset{M\times 1}{\frac{\partial E}{\partial \bar{w}}}$$

$$= \underset{M\times N}{\Phi^T} \underset{N\times M}{\Phi} \underset{M\times 1}{\bar{w}} - \underset{M\times N}{\Phi^T} \underset{N\times 1}{\bar{t}} = 0$$

Recall : $\phi_m(x_n) = \exp\left(-\frac{(x_n - \mu_m)^2}{2s^2}\right)$

$$\bar{w} = (\Phi^T \Phi)^{-1} \Phi^T \bar{t}$$

→ This basis functions based equation is important to understand later topics such as Gaussian Processes and SVM. (Later, it is known as kernel)

(Read sections 3.1.1 and 3.1.2 of the textbook)